Journal of Experimental Child Psychology 194 (2020) 104821

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

The influence of environmental reliability in the marshmallow task: An extension study

Lillie Moffett [a,*], Carol Flannagan [b], Priti Shah [a]

[a] Department of Psychology, University of Michigan, Ann Arbor, MI 48109, USA
[b] University of Michigan Transportation Research Institute, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 15 July 2019
Revised 28 January 2020
Available online 10 March 2020

Keywords:
Delay of gratification
Marshmallow Task
Bayesian statistics
Decision-making
Replication study
Preschool

ABSTRACT

This study is an extension of an experiment where the reliability of children's environment was manipulated before children completed the Marshmallow Task (Cognition, 2013, Vol. 126, pp. 109–114). In that experiment, Kidd, Palmeri, and Aslin found a significant difference in waiting time between two conditions in which the experimenter demonstrated reliability (by returning with promised reward) or unreliability (by not returning with rewardP). Children who had an unreliable experimenter did not wait as long during the Marshmallow Task, suggesting that delay gratification performance may be, in part, based on a rational decision. Due to the important theoretical and practical implications of this finding, we repeated the procedure of this experiment with 60 3- to 5-year-old children (twice as many as in the original study), but in a more familiar context (e.g., children's school instead of a lab). Using Bayesian analyses, we found an effect (albeit smaller than in the original study) of experimenter reliability as well as a significant gender by condition interaction effect.

© 2020 Elsevier Inc. All rights reserved.

Introduction

It is a well-known phenomenon that young children have difficulty in waiting and will often settle for a less desirable reward in the present despite knowledge of a more desirable reward in the future.

* Corresponding author.
E-mail address: lillmoff@umich.edu (L. Moffett).

https://doi.org/10.1016/j.jecp.2020.104821
0022-0965/© 2020 Elsevier Inc. All rights reserved.

The capacity to delay gratification is commonly assessed by the classic Marshmallow Task (Mischel & Metzner, 1962). Preschoolers' ability to wait for the promise of two marshmallows—when one marshmallow lies right under their nose—is associated with a host of later positive life outcomes (Ayduk et al., 2000; Mischel et al., 2011; Schlam, Wilson, Shoda, Mischel, & Ayduk, 2013; Shoda, Mischel, & Peake, 1990) that, until recently, have most often been attributed to children's underlying capacity for self-control (Duckworth, Tsukayama, & Kirby, 2013).

Whereas some consider individual differences in delay of gratification to be driven by more general intrinsic self-control and executive functioning capacities, others have found that individual differences in delay performance are highly influenced by children's environment, suggesting that children may be engaging in a rational decision-making process during the task. This latter interpretation has been less explored, but given the recent evidence illustrating the salient effect of socioeconomic status (SES) on children's delay performance (e.g., Watts, Duncan, & Quan, 2018), it is important to examine processes by which environmental factors may play a role. For example, we know that children from lower-SES backgrounds experience more instability and stress in their environment than do children from higher-SES households (Evans & English, 2002; Frankenhuis, Panchanathan, & Nettle, 2016). It could be that children learn to be sensitive to the provisions of their environment and explicitly *decide* not to wait when realizing that waiting for a reward might not pay off. This interpretation is further supported by evidence from studies illustrating how manipulating the reliability of the environment (Kidd, Palmeri, & Aslin, 2013), manipulating the perceived trustworthiness of an experimenter (Michaelson & Munakata, 2016), and manipulating social trust more generally (Ma, Chen, Xu, Lee, & Heyman, 2018; Mahrer, 1956; Mischel, 1958) all influence the amount of time children wait during a delay of gratification task. Similarly, the reward itself may be evaluated and determined worthy (or not) of waiting (Garon, Longard, Bryson, & Moore, 2012; Liu, Gonzalez, & Warneken, 2018). Thus, there is substantial evidence illustrating that external influences have an effect on children's capacity (or decision) to delay gratification.

One important study that examined the effect of context during a delay of gratification task was carried out by Kidd et al. (2013). In their experiment, children aged 3–5 years were randomly assigned to either a *reliable experimenter* condition or an *unreliable experimenter* condition. At the beginning of the experiment, children were taken to an art project room and provided with materials for a "decorate a cup" activity with older and unappealing materials. Twice, they were offered a choice of using the older/less appealing materials (crayons and then stickers) or waiting for new/better materials. Children waited, in part, because the wait time was relatively short and the older materials were hard to access (e.g., tight lid on crayons). In the reliable condition the experimenter came back with the more preferred materials, whereas in the unreliable condition the experimenter returned empty-handed. Following this art time, participants then participated in the standard Marshmallow Task. Reliability of the experimenter during the art project time had a substantial impact on delay of gratification; children in the reliable group waited an average of 12.03 min, and children in the unreliable group waited only 3.02 min (out of 15 min in total). This dramatic impact of environmental reliability on delay of gratification is noteworthy and suggests that not delaying gratification might be due not to a lack of self-control or impulse control but rather to a deliberate and rational choice (Kidd et al., 2013).

The findings from Kidd et al. (2013) study are a major contribution to the field, but considering that their sample size was relatively small (28 children; 14 per condition) and the effect size of their manipulation was large (d = 1.92), it would be valuable to extend this experiment with a larger sample to examine the robustness of this result (Duncan, Engel, Claessens, & Dowsett, 2014; Gelman & Weakliem, 2009). With a larger sample size, gender differences in this phenomenon can also be analyzed given that there is past work illustrating how boys choose to wait for a better reward (rather than settle for a less desirable immediate reward) more often than do girls (Garon et al., 2012). Furthermore, Kidd et al. (2013) study was conducted in an unfamiliar lab setting, but we do not know whether children tested in a more familiar setting would still be as sensitive to the reliability of a person. Given the historical importance of the Marshmallow Task and the fact that Kidd et al.'s findings have important implications for understanding how and why children delay gratification (or not), we conducted an extension of this study where the unfamiliar experimenter and procedure remained the same, but children were tested in a room at their school instead of in a lab setting. Although this

context may generate different expectations regarding the reliability of the environment, it would be valuable to investigate whether the phenomenon extends to a less novel context. This would then provide insight into whether general trust in the reliability of a context (e.g., school) can buffer the influence of experimenter- or person-specific reliability.

In the current study, data were analyzed using the same analyses that Kidd et al. (2013) used. In addition, we tested whether there were differences by gender and also used a Bayesian approach that combines the results of this study and Kidd et al.'s study. Compared with a more frequentist (classical) paradigm that tests findings against a null hypothesis, Bayesian statistics are useful because they allow us to take into account the prior experimental results and integrate the findings across both studies to draw a more robust conclusion about the similarity of the experimenter reliability effect in a different context (e.g., extension study).

Method

Participants

Institutional review board approval was granted from the University of Michigan. A power analysis was conducted with Kidd et al. (2013) sample to determine the sample size for the current study. Given their effect size, a sample size of 22 children was needed to reach 80% power (Kidd et al.'s study had 28 children). However, because the results from the original study may be larger than those in the more familiar context of the study, we chose a sample size of 60 participants, which can detect an effect size of 0.72 with 80% power and assuming the original study's variance estimate. A total of 61 children were recruited from preschool and kindergarten for this study (M_{age} = 4;6 [years;months], *SD* = 0;8, range = 3;2–5;7). One child was dropped from the sample due to a disruption during the first part of the experiment. Participants of each gender were assigned to condition in an alternating fashion at each site/classroom, leading to an age and gender balance across conditions. This study was conducted at three separate schools in primarily middle- to upper-middle-class neighborhoods in the midwestern region of the United States. Thirty-seven percent of the participants were White (n = 22), 43% were of Middle Eastern descent (n = 26), 10% were of East Asian descent (n = 6), and 10% were reported by their parents to be of mixed races (n = 6). After parental consent was collected, children were individually tested in a separate space from their classroom and all sessions were video-recorded. A single experimenter tested all children in the sample, and no child was previously familiar with the experimenter. Due to the nature of the procedure, the experimenter was not blind to condition.

Procedure

The procedure exactly replicated that of Kidd et al. (2013) study (e.g., materials, wait time, unfamiliar female experimenter) except that children were tested in a room at their own school (a familiar environment) instead of in an unfamiliar lab setting.

Art project task

The entire procedure exactly replicated that of Kidd et al. (2013) study, beginning with an Art Project Task in which children decorated a cup. For this task, children needed to make two choices regarding decorations for their cup. In Choice 1, children were first told that they could either use well-used crayons now or wait for a brand new set of art supplies. In Choice 2, they were then told that they could either use one small sticker now or wait for a new set of fancier stickers. In the reliable condition the experimenter returned with the more desirable choice (new art supplies and new stickers) after children waited for 2.5 min, whereas in the unreliable condition the experimenter failed to return with the more desirable choice. For the exact script used, see Kidd et al. (2013).

Marshmallow task

This part of the procedure also replicated that of Kidd et al. (2013) study. The experimenter told children that they could either eat one marshmallow now or wait for two marshmallows when the

experimenter returned. The experimenter then left the room and returned after 15 min or until children tasted (licked or bit) the marshmallow.

Coding

One coder, blind to condition, recorded when children first tasted (licked or bit) the marshmallow. For reliability purposes, a second coder then coded a randomly selected 15 videos (25%). This second coder was within 2 s of agreement on 13 of the 15 videos (87%), which demonstrates good reliability. The videos that experienced disagreement were instances where a sniff of the marshmallow was mistaken for a lick or it was ambiguous whether a child who picked off a bit of the marshmallow actually ate it. Discrepancies were resolved after discussion between the two coders.

Results

Children in the reliable condition waited an average of 9.58 min (SD = 5.02), whereas children in the unreliable condition waited an average of 7.01 min (SD = 5.22). This reveals a time difference of 2.57 min, which is an effect size (Cohen's d) of 0.49. The original study's effect size was 1.92.

Initial (preregistered) analyses

Original analyses directly mirrored those of Kidd et al. (2013) study and were registered on Open Science Framework (https://osf.io/q7huk/register/565fb3678c5e4a66b5582f67). We conducted an exact Wilcoxon rank-sum test, and results revealed a nonsignificant difference between conditions (W = 572, p = .069). A binary analysis was performed to determine whether condition differences emerged on waiting the full 15 min without tasting the marshmallow. In the reliable condition 9 children (30%) waited the full 15 min, and in the unreliable condition 6 children (20%) waited. A two-sample test for equality of proportions revealed an insignificant difference (χ^2 = 0.80, p = .37) between the two conditions on waiting the full 15 min. Finally, a linear regression was conducted with age, gender, condition, and two interaction terms for age by condition and gender by condition regressed on total time waited. Results are illustrated in Table 1. Gender and the interaction between gender and condition were marginally significant predictors in the model. Boys waited an average of 11.0 min in the reliable condition and 5.8 min in the unreliable condition. Girls, on the other hand, waited an average of 8.2 min in both conditions (see Fig. 1).

Bayesian approach

Bayesian analyses were conducted on both studies. Bayesian methods are particularly appropriate for replication and extension studies because they provide a quantitative way to model *both* effects from the original *and* replication/extension studies. For more information about the advantages of Bayesian statistics in developmental psychology research, see van de Schoot et al. (2014).

First, we analyzed the two studies separately. For each study, we used a truncated normal model while accounting for censoring. The truncated normal was required because measured time cannot be less than zero, and many observed times were fairly close to zero. Censoring occurred at 15 min

Table 1
Linear regression with time waited for marshmallow as the outcome.

	Estimate	Standard error	t Value	p Value
Condition	−9.12	9.37	−0.97	.34
Age	0.01	0.28	0.03	.98
Gender	−7.74	4.59	−1.69	.09
Condition * Gender	5.11	2.84	1.80	.07
Condition * Age	−0.02	0.17	−0.13	.90

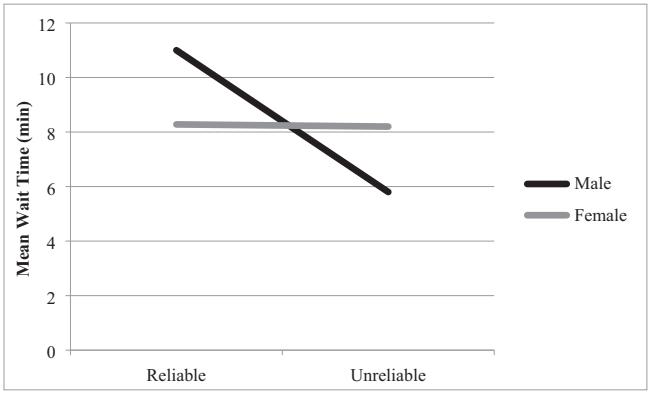

Fig. 1. Illustration of the gender by condition interaction effect. Boys were more sensitive to the reliability of the experiment than were girls.

for any participants who were able to wait until the experimenter returned. Our Stan code is provided in the Appendix.

We initially ran the truncated normal model for each study using only the group membership as a predictor (i.e., analogous to the Wilcoxon test or a *t* test) and allowing separate variance parameters for each group. For each dataset, we generated posterior predicted values for both censored and uncensored observations, and we calculated the group means (grouped by condition) of each of the posterior samples as well as the difference between these.

We ran 20,000 Hamiltonian Monte Carlo (HMC) samples with two chains. HMC diagnostics showed good convergence (Gelman–Rubin $R_{\text{hat}} = 1$). Initially, we ran each model using weakly informative priors on the parameters. For the intercepts (equivalent to the mean for the reliable condition), we used normal (0, 15). This allowed the reliable condition mean to be anywhere in the measured range with reasonable prior probability. For the difference between the unreliable and reliable conditions, we used normal (0, 10), which provides a weak constraint on the difference parameter. For variance parameters, we used half normal (0, 10). The model is described in the equations below:

$y_R, y_U > 0$	$\mu_R \sim N(0,15)$
$y_R \sim N(\mu_R, \sigma_R)$	$\delta \sim N(0,10)$
$y_U \sim N(\mu_R + \delta, \sigma_R)$	$\sigma_R, \sigma_U \sim$ Half Normal (0, 10)

The posterior predictive distribution showed a good match to the data, especially in the extension study. An example of nine overlaid densities can be seen in Fig. 2. Note that the right tails of the densities are expected to be to the right of the sample because they account for censoring. However, those tails cover plausible values of how long children might be able to wait.

A more comprehensive way to look at the posteriors from the two studies is shown in Fig. 3. The two graphs show the distribution of the posterior sample group means across the 20,000 posterior samples for the truncated normal fits to the original (left) and extension (right) data. The red (original) and green (extension) dots show the observed sample mean pairs relative to the fitted distributions for each study. The uncertainty in parameters is much greater in the original study (with lower n) and especially for the reliable condition, in which a large percentage of the original study participants were censored at 15 min (e.g., experimenter returned after 15 min even if children had not touched marshmallow).

Table 2 shows the mean, median, and credible intervals for the difference between posterior sample group means for both models and both datasets. We present this statistic rather than the difference parameter values because the left truncation on the normal distribution means that samples

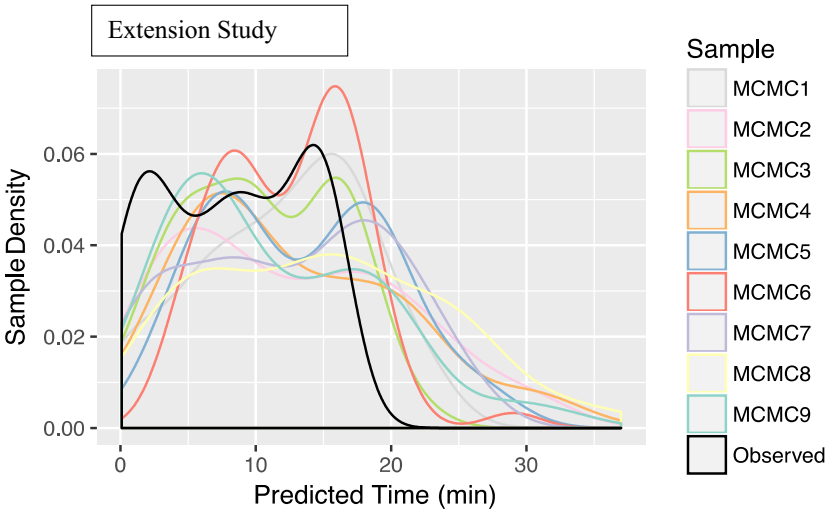

Fig. 2. Density of Markov chain Monte Carlo (MCMC) samples overlaid on observed sample density (black).

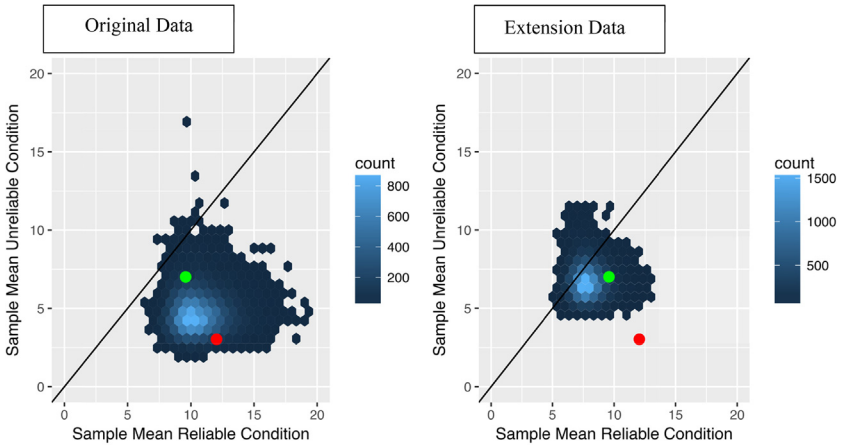

Fig. 3. Distributions of posterior sample means for reliable and unreliable conditions for truncated normal models of original and extension studies. The green dot is the observed sample mean pair for the extension study, and the red dot is the observed sample mean pair for the original study. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Difference in posterior sample means between conditions for the extension and original studies.

Study	Mean	2.5th	50th	97.5th
Extension	−1.13	−3.62	−1.13	1.28
Original	−6.75	−11.92	−6.57	−2.62

Note. Negative values indicate that participants waited longer in the reliable condition.

(i.e., simulated observations) are not symmetrical around the group mean. Conditions with lower mean times will have a greater proportion of their distribution to the right of the mean than will those with higher mean times, reducing the observable differences in the dependent variable. For example,

in the two-group analysis of our extension study, 94% of difference parameters were less than 0, whereas 83% of sample mean differences were less than 0.

We took two approaches to exploring the relationship between the two studies and combining information from them to improve our understanding of the influence of the environmental reliability. Although the studies were conducted in different contexts, the procedures were identical and included an unfamiliar experimenter, so we believe that the two datasets can be combined. First, we used the posterior from the analysis of the original data as a prior for the analysis of the extension study data. Second, we evaluated the extent to which the two samples themselves come from the same underlying distributions (one for each condition).

In the first combined analysis, we used the posterior from the two-group truncated normal analysis of the original data as a prior for analysis of the extension study data. For each of the four parameters (μ_R, δ, σ_R, and σ_U), we used the Markov chain Monte Carlo sample mean and standard deviation of each parameter as parameters in normal and half normal (for sigma parameters) priors. These values, the values for the extension study with weak priors, and the resulting combined estimates of the four parameters are shown in Table 3.

The resulting distribution of posterior sample mean differences is shown in Fig. 4. Here, the density from the first analysis with weakly informative priors is compared with the density when the results of the original study are used as a prior for the extension study. Not surprisingly, the information from the original study increases the size of the differences between the posterior predictions for the two groups. The mean difference between posterior sample means was −3.20 (credible interval = −5.65, −0.95). In plain English, taking into account both the original study and the new study, there is evidence for a 3.2-min difference between the reliable and unreliable conditions with a confidence interval that does not overlap with zero.

Finally, we evaluated the extent to which the samples in the two studies were likely to have come from the same distribution. One way to do this is to look at where the observed samples lie in the distributions in Fig. 3. Note that the extension study sample means are in a credible range for both studies, but the original sample means are not credible in analysis of the extension study. Fig. 5 shows the same plot for the analysis of the extension study using the original results as a prior. Even when the original results are taken into account, the observed sample means from the original study are not credible values based on the model of our extension study. In other words, this means that the effect of experimenter reliability in the original study in an unfamiliar lab context is substantially larger than the effect of experimenter reliability in the current study.

The last analysis of our extension study data incorporated predictors into the truncated normal model. We used gender, age, and the gender by condition and age by condition interactions to predict wait time distributions. These were represented as linear models predicting the mean wait time in the reliable and unreliable conditions. Variances were modeled separately by group as before.

The resulting posterior sample group mean differences had a mean of −2.72 (credible interval = −7.26, 1.74). The regression revealed the previously noted gender by condition interaction such that boys exhibited bigger differences in posterior sample means between the reliable and unreliable conditions than did girls. This difference is shown in Fig. 6, which shows the distribution of posterior sample group mean differences for boys and girls (models were evaluated at a fixed age of 4.5 years, the sample mean).

Table 3
Posterior means and standard deviations of model parameters using each dataset alone and combined.

Model	$\mu_R(SD)$ crI	$\delta(SD)$ crI	$\sigma_R(SD)$ crI	$\sigma_U(SD)$ crI
Original only	15.65 (2.72) (10.61, 21.38)	−12.25 (3.00) (−18.39, −6.60)	8.99 (2.09) (5.69, 13.84)	5.02 (1.09) (3.38, 7.60)
Extension only	10.76 (1.40) (8.14, 13.66)	−2.97 (1.91) (−6.75, 0.72)	7.18 (1.28) (5.18, 10.16)	7.03 (1.15) (5.18, 9.69)
Extension with original as prior	13.54 (1.44) (10.87, 16.50)	−6.72 (1.65) (−10.07, −3.59)	8.47 (1.32) (6.18, 11.29)	6.09 (0.66) (4.90, 7.48)

Note. crI, credible interval.

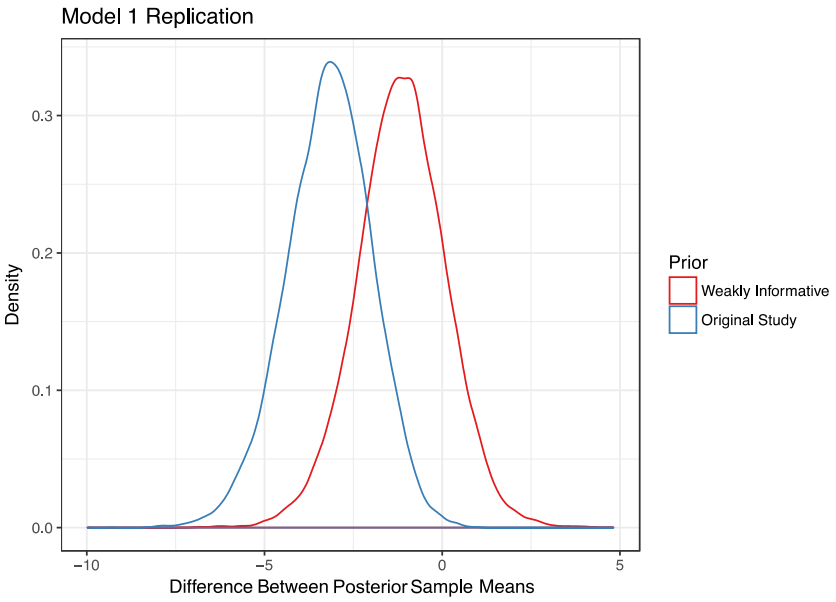

Fig. 4. Difference between condition means in posterior predictive samples drawn from analysis of the extension data under two priors: a weakly informative prior centered at 0 (right curve) and the original study's posterior sample mean (left curve).

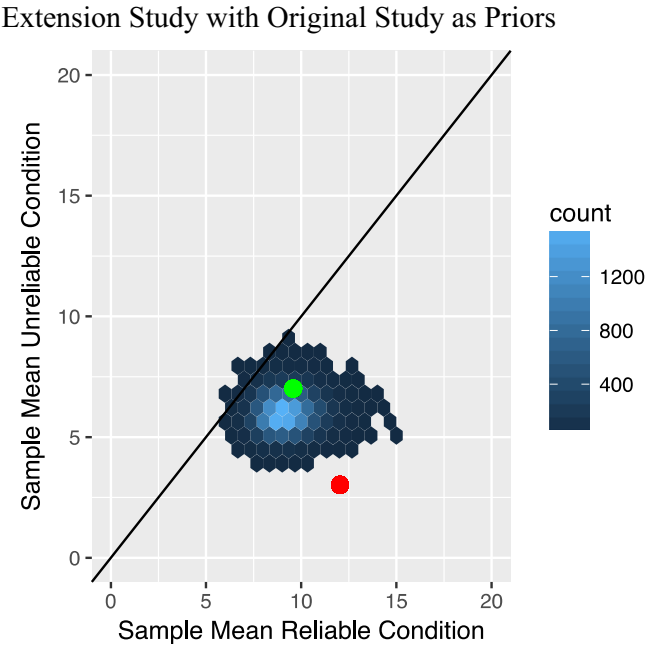

Fig. 5. Bivariate density of sample means by condition for posterior predictive samples drawn from model of extension study data with original posterior means used as priors. The green dot is the observed sample mean pair for the extension study, and the red dot is the observed sample mean pair for the original study. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

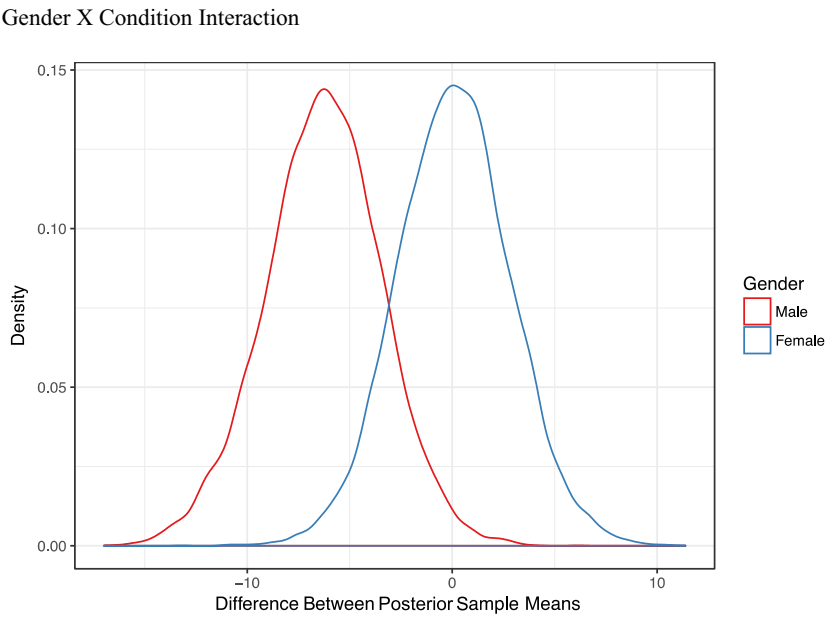

Fig. 6. Distribution of the difference between posterior sample group means separated by boys and girls using the model of the replication data including predictors.

Discussion

The current study extended the work of Kidd et al. (2013), who examined how environmental reliability influenced children's performance in the Marshmallow Task. Although there was no effect of environmental reliability when using a frequentist approach, there was an effect of environmental reliability on preschool children's delay of gratification when using a Bayesian approach—replicating the finding of Kidd et al., but for children tested in a more familiar testing environment. Specifically, when the results of Kidd et al.'s study were used as priors for a Bayesian analysis of our study, the 95% credible interval did not include zero. However, the effect size was less than that in the original experiment, which was conducted in an unfamiliar environment with a smaller sample. For instance, the estimated influence of experimenter reliability on waiting times was much smaller in the extension study ($M = -1.13$ min) compared with the original study ($M = -6.75$ min), and the extension study using the original as prior was in between ($M = -3.20$ min). It is also important to note that the observed pair of sample means from our extension study is credible under all three models (original data only, extension study data only, and extension study using original as prior).

As discussed previously, the environment in the current study was more familiar than the novel lab setting that Kidd et al. (2013) used. It is likely that children are less sensitive to the perceived reliability of the experimenter when they are more familiar with the setting (i.e., school) and, as a result, are more likely to inherently trust experiences in this context. Past studies have illustrated that having increased trust leads to longer waiting in the Marshmallow Task (Ma et al., 2018; Michaelson & Munakata, 2016). The school context is also a place where children are subject to more rules and routines, so it is possible that they are more willing to abide by the experimenter's "rule" to not eat the marshmallow for fear of school punishment or by some implicit "trust" in the reliability of events and routines that occur in school. This raises the theoretical question of whether children are considering the reliability of a person (e.g., experimenter) or the environment itself (e.g., school). Kidd et al. (2013) concluded that children in their study were considering the "environmental reliability" of the

situation; however, these authors did not explicitly tease apart "person" and "environment"; a person is only *part* of one's external environment. Other studies have also illustrated this person-specific sensitivity when the testing setting itself was unchanged and typically in a lab setting (Ma et al., 2018; Michaelson & Munakata, 2016). But it is probable that the perceived reliability of one's overall environment (e.g., home, school, familiar/unfamiliar place) partially overrides an unreliable person present in this environment. For example, family instability, as measured by heightened economic risk, is associated with poor performance in delay of gratification tasks (Sturge-Apple, Davies, Cicchetti, & Hentages, 2017). Future studies should aim to tease apart these aspects of children's environment because it may inform which factors children are more sensitive to when they are developing the capacity to delay gratification.

Our finding that boys were more sensitive to the experimental manipulation than were girls is an important difference from Kidd et al. (2013) study. It is perhaps unsurprising that Kidd et al. did not find an effect of gender because they had nearly twice as many boys as girls in each condition (9 boys and 5 girls) and had a smaller sample size, raising the question of whether boys were driving their main effect in the first place. In the current study, it is certainly puzzling why boys did demonstrate this sensitivity to experimenter reliability. One possible explanation is that boys are not as sensitive to the larger environmental influence (e.g., the school setting) and are less likely to feel pressure to comply with the environmental "rules" and wait for the marshmallow. Past studies have also illustrated that boys are more sensitive to receiving rewards and less sensitive to loss (Fullana, Caseras, & Torrubia, 2003; Garon & Moore, 2007), which might explain boys' increased wait time in the reliable condition (reward) but not their decreased wait time in the unreliable condition (loss). More generally, this gender interaction may also be explained by differential susceptibility theory, which notes that children's genetic makeup makes some individuals (more often boys) more susceptible to environmental influences than others (Belsky & Pluess, 2009). Unlike gender, we did not find an effect of experimental manipulation on performance by child age, which is similar to Kidd et al. (2013) and other studies examining environmental influences on delay of gratification performance (e.g., Michaelson & Munakata, 2016). This is an interesting finding because we did have a wide age range, and delay of gratification and related constructs—such as effortful control—typically undergo large improvements in children between 2 and 4 years of age (Kochanska, Murray, & Harlan, 2000). Future work should perhaps explore why gender—and not age—is more sensitive to environmental reliability and whether children become sensitive to the environment when delaying gratification before 3 years of age.

Conclusion

The current study found that preschool children waited longer in the Marshmallow Task when tested by a reliable experimenter than when tested by an unreliable experimenter (extension of Kidd et al.'s [2013] study). However, this finding appears to be less pronounced when children are tested in a more familiar environment. Furthermore, this sensitivity to experimenter reliability was stronger for boys than for girls. The Bayesian approach that we took may serve as a model for future extension studies in the developmental sciences. This approach allows one to consider outcomes of a new study taking into account the prior data of the replicated study and, subsequently, to compare the relative effect sizes in different circumstances. The results of this study add to the growing body of evidence that performance on delay of gratification tasks, in the short term, are influenced by environmental factors.

Acknowledgments

We thank the participants and their families for contributing their time to this study. The first author was funded by an Institute of Education Sciences fellowship award (R305B150012).

Appendix A. Appendix

Stan code for truncated normal models with censoring and predictors

```
//The normal_lb_rng function is needed to make the posterior predictive draws
//     based on the HMC samples. Stan does not have this built in.

functions {
  real normal_lb_rng(real mu, real sigma, real lb) {
    real p = normal_cdf(lb, mu, sigma);  // cdf for bounds
    real u = uniform_rng(p, 1);
    return (sigma * inv_Phi(u)) + mu;  // inverse cdf for value
  }
}

data {
  int<lower=0> N;           //total number of uncensored observations in all groups
  int<lower=0> N_cens;  //Number of censored values
  int<lower=0> K;          //number of predictors

  vector<lower=0>[N] y_obs;  //dependent measure for all uncensored observations
  vector<lower=0, upper=1>[N]  groupindex; //group index for each
                                    // uncensored case = 0 or 1
  vector<lower=0, upper=1>[N_cens]  groupindex_cens;    //group index for each
                                    // censored case = 0 or 1
  matrix[N,K] x;     //design matrix of predictors for uncensored cases (N cases, K
                                    //parameters)
  matrix[N_cens,K] x_cens; //design matrix of predictors for
                                    //  censored cases
  vector[2] alpha_prior_pars; //prior normal variance on all mu parameters
  real<lower=0> sigma_prior_var; //prior normal variance on all sigma parameters
  vector[K] beta_mu_prior; //prior normal mean on all mu parameters
  vector<lower=0>[K] beta_sig_prior; //prior normal variance on all mu parameters
                                    // real beta_mu_prior;
                                     // real<lower=0> beta_sig_prior;
  real<lower=max(y_obs)> U;  //value at which observations are censored
                                    // (in this case, 15)
}

parameters {
  real alpha;                       //intercept
  real<lower=0> sigma0;        //group 0 standard deviation
  vector[K] beta;                   //coefficients on predictors
  real<lower=0> sigma1;      //group 1 standard deviation
  real<lower=U> y_cens[N_cens];  //imputes y values for censored observations
```

```
  }


model {

  alpha ~ normal(alpha_prior_pars[1],alpha_prior_pars[2]);  // mu's can be negative
  sigma0 ~ normal(0,sigma_prior_var);  //sigma & mu prior vars set separately
  beta ~ normal(beta_mu_prior,beta_sig_prior);  //diff prior set separately
  sigma1 ~ normal(0,sigma_prior_var);


  for (i in 1:N ) {
    y_obs[i] ~ normal(alpha+x[i]*beta,
          sigma0 * (1-groupindex[i]) + sigma1 * groupindex[i]);
  }

  for (i in 1:N_cens) {
y_cens[i] ~ normal(alpha+x_cens[i]*beta,  sigma0 * (1-groupindex_cens[i]) + sigma1 *
groupindex_cens[i] );
          //updates likelihood for censored obs
  }


}

generated quantities {
  vector <lower=0>[N] y_rep;
  real samplemean_gp0;
  real samplemean_gp1;
  real samplemeandiffc2c1;
  real mu_gen;

  samplemean_gp0=0;
  samplemean_gp1=0;

  for(n in 1:N) {

    mu_gen=alpha+x[n]*beta;

// This loop generates quantities for the uncensored observations
//        Note that the censored observations are imputed above because
//        they are treated as parameters
   y_rep[n] = normal_lb_rng(mu_gen,
      (sigma0 * (1-groupindex[n]) + sigma1 * groupindex[n]),0);
//update sample means for group 0 and 1 with new post pred draw
```

```
    samplemean_gp0=samplemean_gp0+y_rep[n]*(1-groupindex[n]);
    samplemean_gp1=samplemean_gp1+y_rep[n]*groupindex[n];
    }


//update sample means with censored obs (imputed previously)

  for (n in 1:N_cens) {
    samplemean_gp0=samplemean_gp0+(y_cens[n]*(1-groupindex_cens[n]));
     samplemean_gp1=samplemean_gp1+(y_cens[n]*groupindex_cens[n]);
    }

//sum(groupindex)+sum(groupindex_cens) gives the total sample count for gp 1
//quantities below compute the group means and the difference of those means

samplemean_gp0=samplemean_gp0/((N+N_cens)-  (sum(groupindex)+sum(groupindex_cens)));
 samplemean_gp1=samplemean_gp1/(sum(groupindex)+sum(groupindex_cens));
 samplemeandiffc2c1=samplemean_gp1-samplemean_gp0;

}
```

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2020.104821.

References

Ayduk, O., Mendoza-Denton, R., Mischel, W., Downey, G., Peake, P. K., & Rodriguez, M. (2000). Regulating the interpersonal self: Strategic self-regulation for coping with rejection sensitivity. *Journal of Personality and Social Psychology, 79*, 776–792.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychological Bulletin, 135*, 885–908.

Duckworth, A. L., Tsukayama, E., & Kirby, T. A. (2013). Is it really self-control? Examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin, 39*, 843–855.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417–2425.

Evans, G. W., & English, K. (2002). The environment of poverty: Multiple stressor exposure, psychophysiological stress, and socioemotional adjustment. *Child Development, 73*, 1238–1248.

Frankenhuis, W. E., Panchanathan, K., & Nettle, D. (2016). Cognitions in harsh and unpredictable environments. *Current Opinion in Psychology, 7*, 76–80.

Fullana, M. A., Caseras, X., & Torrubia, R. (2003). Psychometric properties of the Personal State Questionnaire in a Catalan sample. *Personality and Individual Differences, 34*, 605–611.

Garon, N. M., Longard, J., Bryson, S. E., & Moore, C. (2012). Making decisions about now and later: Development of future-oriented self-control. *Cognitive Development, 27*, 314–322.

Garon, N., & Moore, C. (2007). Developmental and gender differences in future-oriented decision-making during the preschool period. *Child Neuropsychology, 13*, 46–63.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power. *American Scientist, 97*, 310–316.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environment reliability. *Cognition, 126*, 109–114.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*, 220–232.

Liu, S., Gonzalez, G., & Warneken, F. (2018). Worth the wait: Children trade off delay and reward in self- and other-benefiting decisions. *Developmental Science, 6* e12702.

Ma, F., Chen, B., Xu, F., Lee, K., & Heyman, G. D. (2018). Generalized trust predicts young children's willingness to delay gratification. *Journal of Experimental Child Psychology, 169*, 118–125.

Mahrer, A. R. (1956). The role of expectancy in delayed reinforcement. *Journal of Experimental Psychology, 52*, 101–106.

Michaelson, L. E., & Munakata, Y. (2016). Trust matters: Seeing how an adult treats another person influences preschoolers' willingness to delay gratification. *Developmental Science, 19*, 1011–1019.

Mischel, W. (1958). Preference for delayed reinforcement: An experimental study of cultural observation. *Journal of Abnormal and Social Psychology, 56*, 57–61.
Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., ... Shoda, Y. (2011). "Willpower" over the lifespan: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience, 6*, 252–256.
Mischel, W., & Metzner, R. (1962). Preference for delayed reward as a function of age, intelligence, and length of delay interval. *Journal of Abnormal and Social Psychology, 64*, 425–431.
Schlam, T. R., Wilson, N. L., Shoda, Y., Mischel, W., & Ayduk, O. (2013). Preschoolers' delay of gratification predicts their body mass 30 years later. *Journal of Pediatrics, 162*, 90–93.
Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology, 26*, 978–986.
Sturge-Apple, M. L., Davies, P. T., Cicchetti, D., & Hentages, R. T. (2017). Family instability and children's effortful control in the context of poverty: Sometimes a bird in the hand is worth two in the bush. *Development and Psychopathology, 29*, 685–696.
van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*, 842–860.
Watts, T., Duncan, G. J., & Quan, H. (2018). Revisiting the Marshmallow Test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science, 29*, 1–19.